Emergency Medicine Journal

Predicting need for hospital admission in patients with traumatic brain injury or skull fractures identified on CT imaging: a machine learning approach.

Journal:	*Emergency Medicine Journal*
Manuscript ID	emermed-2020-210776.R2
Article Type:	Original research
Date Submitted by the Author:	27-Feb-2021
Complete List of Authors:	Marincowitz, Carl; The University of Sheffield, School of Health and Related Research (ScHARR) Paton, Lewis; University of York Alcuin College Lecky, Fiona; University of Sheffield, School of Health and Related Research Tiffin, Paul; University of York Alcuin College
Keywords:	Trauma, research, Trauma, head, imaging, CT/MRI

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Predicting need for hospital admission in patients with traumatic brain injury or skull fractures identified on CT imaging: a machine learning approach.

Carl Marincowitz[1] NIHR Clinical Lecturer Emergency Medicine, MB BChir, PhD, MSc, BA (Hons), MRCEM

Lewis W. Paton[2] Research Fellow, BSc (Hons), PhD

Fiona E. Lecky[3] Professor, Honorary Emergency Medicine Consultant, MB ChB, FRCS, DA, MSc, PhD, FRCEM

Paul A. Tiffin[4] Reader, BMedSci (Hons), MBBS, FRCPsych, MD

1. **Corresponding Author.** Centre for Urgent and Emergency Care Research (CURE), Health Services Research School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK, Fax: +44 (0)114 222 0749 Tel: (+44) (0)114 222 4345, Email: c.marincowitz@sheffield.ac.uk

2. Department of Health Sciences, University of York, Alcuin Research Resource Centre, Heslington, York, YO10 5DD, Tel +44 (0) 1904 321516, Fax: +44 (0) 1904 32 3433, Email: lewis.paton@york.ac.uk

3. Centre for Urgent and Emergency Care Research (CURE), Health Services Research School of Health and Related Research, University of Sheffield, Regent's Court Regent Street, Sheffield, S1 4DA, +44 (0)114 2220834, @CURE_SCHARR, Email: f.e.lecky@sheffield.ac.uk

4. Hull York Medical School York and Department of Health Sciences, University of York, Alcuin Research Resource Centre, Heslington, York, YO10 5DD, Tel +44 (0) 1904 321516, Fax: +44 (0) 1904 321117, Email: paul.tiffin@york.ac.uk

Abstract

Background

Patients with mild traumatic brain injury (TBI) on CT scan are routinely admitted for inpatient observation. Only a small proportion of patients require clinical intervention. We recently developed a decision rule using traditional statistical techniques that found neurologically intact patients with isolated simple skull fractures or single bleeds <5mm with no pre-injury anti-platelet or anti-coagulant use may be safely discharged from the ED. The decision rule achieved a sensitivity of 99.5% (95% CI, 98.1–99.9) and specificity of 7.4% (95% CI, 6.0–9.1) to clinical deterioration. We aimed to transparently report a machine learning approach to assess if predictive accuracy could be improved.

Methods

We used data from the same retrospective cohort of 1699 initial Glasgow Coma Scale 13–15 patients with injuries identified by CT who presented to three English Major Trauma Centres between 2010-2017 as in our original study. We assessed the ability of machine learning to predict the same composite outcome measure of deterioration (indicating need for hospital admission). Predictive models were built using gradient boosted decision trees which consisted of an ensemble of decision trees to optimise model performance.

Results

The final algorithm reported a mean Positive Predictive Value of 29%, mean Negative Predictive Value of 94%, mean AUC (C-statistic) of 0.75, mean sensitivity of 99% and mean

specificity of 7%. As with logistic regression, GCS, severity and number of brain injuries were found to be important predictors of deterioration.

Conclusion

We found no clear advantages over the traditional prediction methods, although the models were, effectively, developed using a smaller data set, due to the need to divide it into training, calibration and validation sets. Future research should focus on developing models that provide clear advantages over existing classical techniques in predicting outcomes in this population.

What is already known on this subject

We have previously empirically derived a clinical decision rule to select low risk patients with injuries on CT imaging following head trauma for discharge from the ED using traditional statistical methods, based on logistic regression. The decision rule is highly sensitive but lacks specificity and implementation would allow only a small proportion of patients to be discharged. Machine learning may theoretically improve the accuracy of prediction, allowing more patients to be safely discharged.

What this study adds

Using data from the same cohort as our previous study we used a machine learning approach to predict which patients in the sample were likely to deteriorate. We found no clear improvement in prediction over a model previously developed using a classical statistical approach.

Key Words:

Mild Traumatic Brain Injury; Prognostic modelling; Machine Learning; Intra-cranial haemorrhage; Minor Head Injury.

Introduction

There are 1.4 million annual attendances to Emergency Department (ED) in England and Wales following head trauma.[1] Of these, 95% of patients attend with an initial Glasgow Coma Scale (GCS) score in the range 13-15 and are defined as having a minor head injury.[2] Around 7% of these patients have brain injuries and skull fractures identified by CT Imaging.[3] In the UK, patients with injuries identified by CT are routinely admitted for observation, although only a small proportion clinically deteriorate.[4] Internationally, some advocate routine admission of patients with injuries on CT to higher dependency areas due to the risk of deterioration, whilst other advocate use of the Brain Injury Guideline (BIG) criteria to select patients for discharge from the ED.[5 6]

Accurate risk prediction of clinically important deterioration in GCS 13-15 patients with traumatic injuries identified by CT imaging could allow the discharge of low risk patients from the ED. Patients with expanding intra-cranial haemorrhage can rapidly and catastrophically deteriorate. This risk must be weighed against the potential advantage of any reduction in hospital admissions. Thus, the use of predictive models to select patients for discharge may be controversial in some clinical settings. The consequences of discharging a patient who deteriorates (a 'false negative') are much greater than admitting a patient who remains stable (a 'false positive'). Therefore, accurate prediction of patients who will not deteriorate is more useful than accurately predicting every patients' risk of deterioration. We recently developed a risk prediction model and decision rule for discharge from the ED for this TBI population using traditional statistical approaches.[7 8] Our derived decision rule outperformed existing guidelines, achieving a high sensitivity to a composite outcome of deterioration encompassing need for hospital admission, but lacked specificity.

Logistic regression, using maximal likelihood estimation, optimises predictive accuracy across the range of possible probabilities of deterioration. Advocates of machine learning have highlighted that more flexible modelling techniques may better capture non-linear relationships and interactions between the variables in the data. The use of 'ensemble

learning', which combines the results of multiple models to make final predictions, is a way of addressing the 'bias-variance trade-off'. That is, the potential bias from multiple models can be averaged out, or otherwise combined, to achieve more consistent predictive accuracy. Thus, theoretically, machine learning based prediction could achieve higher levels of accuracy compared to traditional statistical modelling approaches.

However, at least for structured data (i.e. that already in numeric format) this has not been firmly established. A recent systematic review reported that, on average, machine learning-based models tended to outperform predictive models that use logistic regression techniques, but only for studies deemed at high risk of bias.[9] Moreover, others have raised concerns that machine learning derived models are prone to 'overfitting'. That is, they replicate the relationships in the data being used to train them accurately but may fail to generalise accurately to new, unseen data sets. There have also been concerns over a lack of transparency and consistency in reporting the results from observational studies using machine learning.[10] This raises issues with the validity and generalisability of the results reported from machine learning studies that purport to form the basis of current or future clinical decision support tools.

We therefore aimed to use machine learning to develop a predictive model which can accurately identify patients with TBI and skull fractures on CT imaging at very low risk of deterioration who could be safely discharged. We used the same data set as in Marincowitz et al.[7 8] so that we were able to understand the predictive potential of machine learning, compared to the tool developed using traditional statistical approaches. Our objective was to build a machine learning model and report our results in a way which was both transparent, reproducible and accurately quantified uncertainty around the predictive precision. By doing so we aimed to address previous criticisms and establish whether the potential advantages of such an approach, employing the latest methods to machine learning using structured data, outweighed any limitations inherent to the approach in this context.[10]

Materials and methods

Study Design

Data were analysed from an existing retrospective cohort study using case note review of TBI patients presenting to the ED between 2010-2017 at three Major Trauma Centres in England: Hull University Teaching Hospital NHS Trust, Salford Royal NHS Foundation Trust, and Addenbrooke's Hospital (Cambridge University Hospitals NHS Foundation Trust). Both a detailed study protocol[7] and a cohort study using traditional statistical techniques[8] have previously been published. We previously used multivariable logistic regression with bootstrap internal validation to derive a predictive model which included: initial GCS, pre-injury anti-platelet or anti-coagulant use, first neurological examination, number of injuries on CT imaging, severity of brain injury, severity of extra-cranial injuries and initial haemoglobin value. Our previously derived model is presented in Supplementary Material 1.

Inclusion Criteria

Patients aged ≥16 with a presenting GCS score of 13-15 who attended the ED following acute head trauma and had injuries reported on CT brain scan. The latter was defined as: skull fractures, extradural haemorrhage, subdural haemorrhage with an acute component, traumatic intra-cerebral haemorrhage, contusions, traumatic subarachnoid haemorrhage and traumatic intra-ventricular haemorrhage.

Exclusion Criteria

Patients were excluded where: a non-traumatic cause of intra-cranial haemorrhage was indicated, pre-existing CT abnormality prevented determining whether acute injury had occurred and patients transferred from other hospitals.

Primary Outcome

A composite measure of deterioration aimed at encompassing need for hospital admission was used. This included up to 30 days following ED attendance any of: death attributable to TBI, neurosurgery, seizure, a drop in GCS>1, ICU admission for TBI, intubation or hospital readmission for TBI. Where reason for death, ICU admission or readmission was unknown it was attributed to TBI deterioration.

Data collection

ED CT brain scan requests and reports were screened at each centre to identify patients with traumatic brain injuries or skull fractures. Patients with identified injuries were matched to their full electronic and written case records to determine if they met the inclusion criteria data. Where they did so, data were extracted by trained research staff using a standardised electronic proforma on patient deterioration outcomes and candidate predictors.

Data pre-processing

For each run of model building and testing the data were equally split into three subsets. These formed a 'training set' on which to develop the predictive algorithm, a 'calibration set' to build the model for probability recalibration (see below), and a 'test set' which is 'held back' to validate the final algorithm. Stratified random sampling was used to ensure equal distribution of the primary composite outcome of deterioration between sets.

Predictive model building

Our predictive models were built using gradient boosted decision trees via the CatBoost package [12] in R.[13] Gradient boosted decision tree models consist of an ensemble of decision trees, aiming to optimise model performance. The method was selected as it is known to work well even with small and medium-sized data sets (i.e. several hundred to several thousand observations). This approach combines a number of methodological approaches to prediction; the use of decision trees; 'ensembling'- where numerous slightly differing models are created, and the results averaged or voted on, and; 'boosting' where the algorithm successively focuses on the observations where the outcome is increasingly difficult to predict. By combining all three approaches, gradient boosted trees tend to outperform algorithms which only use one or two of these methods. There is evidence for this in that the majority of winning solutions in the 'Kaggle' prediction competitions feature ensembles of boosted trees.[14] CatBoost extends this approach by the way it treats categorical (and in this case, ordered) predictor variables. The software recodes such categorical variables to numeric, depending on their observed relationship with the outcome of interest. This can potentially increase the amount of information available to predict the outcome of interest. The CatBoost algorithm has two main 'hyperparameters'

that can be changed in order to improve predictive accuracy and generalisability: the number of decision trees to grow, and; the number of variables to select at each node of the trees. The process of choosing hyperparameter settings is known as 'tuning'.

Model building proceeded as follows (also see Figure 1). When predicting relatively uncommon outcomes it is important to stop the algorithm focusing on achieving high accuracy by predicting the most prevalent outcome (in this case, a lack of deterioration). For this reason, we used 'Synthetic Minority Over-sampling Technique' (SMOTE) in order to create synthetic observations with the relatively uncommon outcome of deterioration, in the training data set.[15] These synthetic patients are based on the actual data on patients with recorded deterioration, and are created using a 'K-nearest neighbours' approach to ensure the training data set has an apparent 50:50 split of participants with the two outcome types (deterioration/no deterioration). We used the default value of k=5. Thus, this pre-processing step helps the algorithm to train to predict the less common outcome (in this case, deterioration). The CatBoost model is then fitted to the training data set, learning how to link the predictor variables to the outcomes. This step involves a 'tuning' phase where the model hyperparameters (e.g. number of decision trees) are altered in order to optimise predictive performance. A grid search over possible values of the hyperparameters was performed in order to find the combination of hyperparameters that maximises the area under the receiver operator characteristic curve (AUC- equivalent to the 'C-statistic') on the training data set. This was done on a sample of training data. The final model is then applied to the previously unseen test data set to predict the class (i.e. deterioration or not) and probability of deterioration for each individual in the test data set.

The predicted probabilities from a decision tree classification tend to cluster around a central point in order to maximise the accuracy metric used to optimise the algorithm. This means that the accuracy at predicting one class versus the other is maximised. However, the resulting predicted probabilities tend not to reflect the true underlying probabilities. This matters if, for example, one wishes to change the threshold for the predicted probability for a case and a non-case in order to, say, minimise false negative cases. Predicted probabilities from such machine learning models can be mapped on to those more likely to reflect the true underlying probabilities through a process known as recalibration. This involves building a second model using a separate portion of training data, not previously used for

building the original machine learning model. This second model seeks to predict the true underlying probabilities, as represented approximately by observed frequencies of the outcome type, from those predicted by the first-phase machine learning model. In this case we used an isotonic regression model on this separate subset of data (the calibration set) to link the predicted probabilities from the predictor variables to the approximate probability of observing the actual outcome (deterioration).[16] Thus, running both the machine learning model and the recalibration model in series provided the final predicted probabilities which can be used to classify the patients in the final, unseen, validation, data set in terms of the risk of deterioration (high vs low risk). Metrics of model performance (e.g. AUC, Negative Predictive Value etc) were then calculated.

Due to the stochastic nature of this algorithm development (e.g. data set splitting, imputation etc.) we repeated the entire process 2,500 times, and the performance metrics were stored for each run. The exception to this was the estimation of the optimal model hyperparameters in the 'model tuning' phase, which we performed only once. In this regard tuning was only performed once, on a single training sample, which was itself then split into a tuning training set and a tuning validation set. Performing tuning only once eased computational requirements, which is possible due to the stability of the results generated from the tuning phase. The optimal model hyperparameters from the second iteration onwards are thus set at the values decided in the tuning phase for the first iteration. The overall performance of the models was evaluated by calculating the mean accuracy metrics over the 2500 iterations. A measure of the spread of the results was calculated using the values at the 2.5th percentile and the 97.5th percentile, to give the central 95% interpercentile range.

As the aim of the model was to decide which patients were relatively safe to discharge from the emergency department we selected an overall predicted probability threshold that led to a relatively high negative predictive value (NPV), albeit at the expense of positive predictive value (PPV). That is, we wanted a predictive system that was relatively good at deciding which patients are safe to go home, even if a significant proportion were flagged as requiring further observation, which might be unnecessary. The cost of false positives, in terms of patient care and potential consequences, was lower than that for false negatives. Thus, our aim of recalibration was to achieve a diagnostic prediction system that performed

at least as well as the BIG criteria[8] (i.e. an NPV of at least 96.5% and minimum PPV of 28%). Our use of a separate recalibration model for the initial predicted probabilities allowed us to move the threshold for the predicted probabilities in this way. This had no impact on overall model performance, with a negligible impact on the AUC values for the model, though recalibration reclassified the error types produced. This meant that the final model output could be adjusted to minimise the risk of false negatives (patients predicted to be at low risk, but who actually did deteriorate) whilst maintaining acceptably useful levels of specificity (i.e. the ability to identify 'true negatives').

Missing data

Missing values for predictor variables were imputed using a single imputation via the Amelia II package for R, which uses an Expectation-Maximization Bootstrap based (EBM) algorithm.[11] This process was stochastic, and each iteration of model building included a new round of imputation. Thus, the missing data imputation was incorporated into the loop of data set splitting and model building. This was important as to account for the uncertainty of this process when evaluating the overall performance of the models.

Ethics

NHS Research Ethics Committee Approval was granted by West of Scotland REC 4 reference: 17/WS/0204. As a retrospective case review conducted by members of the direct care team, consent was not requited.

Patient and Public Involvement

The Hull and East Yorkshire NHS Trust Trans-Humber Consumer Research Panel and Hull branch of the Headway charity were consulted in the initial stages of developing the research questions addressed in this study.

Results

Study population

Figure 2 summarises screening of ED CT requests and inclusion of patients on matching to case records at each centre and Table 1 the population characteristics and candidate

variables. The cohort was mostly male, with around half of patients >60 years of age and one quarter with either pre-injury anticoagulant or -platelet use. The cohort was mostly male, with around half of patients aged over 60 and quarter with either pre-injury anti-coagulant or anti-platelet use. 470 patients (27.7%; 95% CI: 25.5% to 29.9%) clinically deteriorated as defined by the primary outcome. A total of 223 patients (13.1%; 95% CI, 11.6–14.8) underwent neurosurgery were admitted to ICU or were intubated. A total of 72 patients had deaths attributable to TBI. A total of 471 patients had data missing from at least one candidate variable on case note review.

Table 1. Variables and population characteristics.

Candidate Variable	Category	Mean (SD), min-max OR N (%)	Missing data (N)
Age	Years	58.2 (SD 23.3) 16-101 Age≥65 = 44.9%	0
Sex	Male Female	67% (Median Age= 52) 33% (Median Age= 69)	0
Glasgow Coma Scale	15 14 13	976 (58%) 533 (31%) 185 (11%)	5 (0.3%)
Mechanism of Injury	Assault Fall RTC Sport Other	228 (13%) 1090 (64%) 298 (18%) 21 (1%) 30 (2%)	31 (1.8%)
Intoxicated	Yes	494 (29%)	38 (2.2%)
Seizure pre-hospital or in ED	Yes	74 (4%)	10 (0.6%)
Vomit pre-hospital or in ED	Yes	310 (18%)	12 (0.7%)
Preinjury Anti-coagulation or anti-platelets	Anticoagulation use Antiplatelet use Both	155 (9%) 294 (17.3%) 8 (0.5%)	0
Abnormal First Neurological Examination	Yes	233 (14.5%)	89 (5.2%)
Initial Blood pressure	Mean Arterial Pressure mmHG	98.5 (SD 17) 43-193	61 (3.6%)
Initial Oxygen Saturation	%	97.4 (SD 2.4) 80-100	59 (3.5%)
Initial Respiratory Rate	RR per Min	17.9 (SD 3.5) 10-48	94 (5.5%)
Haemoglobin	Grams/litre	136 (SD 19.1) 68-265	211 (12.4%)
Platelet Value	10^9/L	232 (SD 77) 2-742	211 (12.4%)
Number of Injuries on CT	1 2 3	824 (48.5%) 400 (23.6%) 217 (12.7%)	0

	4 5 Multiple diffuse injury*	142 (8.4%) 103 (6.1%) 13 (0.8%)	
Injury severity on CT (Modified Marshall Classification described in detail supplementary Material 2)	Isolated Simple Skull Fractures Isolated Complex Skull fractures 1-2 bleeds < 5mm (total) No or minimal mass effect Significant midline shift High/mixed-density lesion Cerebellar/Brain stem injury	66 (3.9%) 123 (7.2%) 208 (12.2%) 1001 (58.9%) 159 (9.4%) 122 (7.2%) 22 (1.2%)	0
Any Skull Fracture (simple)	Yes	316 (19%)	0
Any Skull Fracture (complex)	Yes	360 (21%)	0
Contusion	Yes	580 (34%)	0
Extradural bleed	Yes	135 (8%)	0
Intraparenchymal haemorrhage	Yes	240 (14%)	0
Subdural bleed	Yes	694 (41%)	0
Intra-ventricular bleed	Yes	50 (3%)	0
Subarachnoid bleed	Yes	536 (32%)	0
Rockwood Clinical Frailty Scale (CFS)	Patients under 50 CFS 1-3 CFS 4-6 CFS 6-9	649 (39%) 642 (38%) 308 (18.5%) 72 (4.5%)	28 (1.6%)
Comorbidity	Charlson Index	1.4 (SD 2.9) 0-28 (range)	20 (1.2%)
ISS	Body regions excluding head	5.2 (SD 5.2) 0-75 (range)	0

Model parameters

The final model hyperparameters were determined in the tuning phase. In this respect 200 trees were created at each run, with nine predictor variables selected, at random from the data set, to be used to create a split at each node (branch split).

Model performance

For each of our 2500 model runs, our test data set consisted of a random sample of 576 patients. Of these, the median number of patients the model predicted could be discharged across our 2500 models was 26 ('true negatives'), and the median number of deteriorations of those 'discharged' (i.e. 'false negatives') was one. As can be seen from the results in Table 2, the mean Negative Predictive Value (NPV) indicates that, on average, 94% of those who were recommended for discharge by the model did not deteriorate. Across all 2500 runs of our model, the value of the 2.5th percentile for NPV was 81%, and the 97.5th percentile was 100%. The mean Positive Predictive Value (PPV) of our models was 29% (2.5-97.5th interpercentile range 28% to 31%). The mean sensitivity was 0.99 (0.96 to 1.00) and the

mean specificity 0.07 (0.01 to 0.17). The mean area under the curve (AUC– equivalent to a C-statistic), an overall metric of the potential utility of the model, was 0.75, with the interpercentile range of this value being 0.71 to 0.78.

Table 2. Predictive ability of the machine learning based models in the test (validation) data sets according to mean accuracy metrics. The model was built and tested 2500 times to estimate the 2.5th percentile and 97.5th percentile values for the performance metrics.

Metric	Mean performance (2.5th to 97.5th interpercentile range)
Accuracy	0.32 (0.28 to 0.40)
Area under the curve (AUC)	0.75 (0.71 to 0.78)
Sensitivity ('true positive rate')	0.99 (0.94 to 1.00)
Specificity ('true negative rate')	0.07 (0.00 to 0.19)
Positive Predictive Value	0.29 (0.28 to 0.31)
Negative Predictive Value	0.94 (0.81 to 1.00)

The CatBoost process did not produce interpretable models as such. However, the output for each run of the model produced 'importance' metrics for the predictors. This metric gives a normalised score to each variable which describes how much the prediction changes if the value of the predictor changes. Ranking the predictors by the mean importance scores therefore gives some indication of which variables the model finds most useful in predicting deterioration status. In Table 3 we provide the mean importance scores for the predictors, averaged over 100 runs. As can be seen in Table 3, we observed that severity of the injury is

deemed most important, followed by GCS, number of injuries, the particular hospital the patient was admitted to, and the presence of subdural haemorrhage.

Table 3. Ranked mean 'importance' of the features (predictors) in the model, averaged over 100 runs.

Feature	Mean importance
Injury severity on CT (Modified Marshall Criteria)	22.69
Glasgow Coma Scale	10.85
Number of injuries	9.67
Hospital admitted to	4.62
Subdural bleed	4.18
Comorbidity (Charlson Index)	4.05
Skull fracture type	3.97
Rockwood Clinical Frailty Scale	3.88
Haemoglobin (g/litre)	3.63
Initial blood pressure (Mean Arterial Pressure)	3.34
Age	3.17
Platelet value	3.16
Initial respiratory rate	2.73
Contusion	2.53
Pre-injury anti coagulation or anti-platelet use	2.41
Initial oxygen saturation	2.41
Subarachnoid bleed	2.39
ISS	2.37

Intoxicated	2.36
Sex	1.51
Vomit pre-hospital or in ED	0.85
Intraparenchymal haemorrhage	0.61
Extradural bleed	0.57
Seizure pre-hospital or in ED	0.25
Intra-ventricular bleed	0.17

Discussion

This is the first study to report the performance of a machine learning approach to predicting the need for hospital admission in this TBI population. Our final algorithm, over 2500 runs, reported a mean PPV of 0.29, mean NPV of 0.94, mean AUC (C-statistic) of 0.75, mean sensitivity of 0.99 and mean specificity of 0.07. These performance metrics are broadly the same as those recently reported for a classical approach to predictive modelling on the same data set using logistic regression and the BIG criteria, although we report a slightly lower mean NPV (94%) than both the BIG criteria (96.5%) and the logistic regression model (97.7%).[8]

The modelling process suggested that the most important variables for predicting deterioration were injury severity, GCS and the number of injuries. While a direct comparison with the previous logistic model developed on these data is not possible, due to some differences in data management and sampling (i.e. in the present study the data set was divided into three portions), the largest odds ratios in the logistic model also related to injury severity, GCS and number of injuries.[8] Other predictors in the logistic regression model included extra-cranial ISS value, anti-coagulant and anti-platelet use, an abnormal neurological examination and haemoglobin value. The presence of specific types of injury appeared more important in the machine learning models and this may be due to the modelling being able to account for interactions between injuries when co-occurring.

Strengths and limitations

Our model appears similar to the previous logistic regression model in terms of both performance metrics and those variables apparently most important in predicting deterioration. The sole advantage of using the machine learning approach in this context appeared to be that the model was developed on much fewer data – approximately one third- than the previous one, employing a classical statistical approach. Our study had a sample size powered to derive a predictive model from our candidate variables using multivariable logistic regression for our original study. This, however, represents a relatively small sample size for developing machine learning models. Moreover, the effective sample size in the present study was smaller still because of the requirement to recalibrate the probabilities from the models being developed. Despite this, it managed to achieve broadly similar performance metrics.

Theoretically, given greater data availability, the machine learning model may have outperformed the classical approach. It may be possible to achieve larger effective sample sizes via alternative methodological approaches. We split our data into three equal sets ('training', 'calibration' and 'test). However, this may not be the optimal division of the original data, and this could have been assessed using sensitivity analyses. Within this, it could also be worth, in future studies, considering stratified training sets to account for key predictor variables. It may also have been possible to reduce 'data spend' by using a cross-validation approach to model calibration, rather than having had a third, separate, portion. This would have required the data to be split only into training and test data sets. Training and calibration of the model would then take place on different 'folds' (i.e. further subsets) of these data, rather than using a separate 'calibration' data set as we did in this study. In this study we used a separate data set and isotonic regression as the approach was easily implemented in the workflow. In addition, the relative sparsity of one of the two outcome categories (deterioration) may have meant that the recalibration model may have benefited from a larger number of such outcomes being present in the data set portion it was built on. However, we recognise that alternative methodologies, such as recalibration using cross-

fold validation, may have worked at least as well, or perhaps better, in the context of a relatively small data set.

Machine learning models often overfit to the data on which they are trained. This leads to poorer performance in external, validation data, and hence, impaired generalisability. The 'CatBoost' algorithm used here includes an 'overfitting detector' which can stop the model training process if overfitting is observed during the training process. [17] Our use of previously unseen ('hold out') validation data samples also would have helped to ensure realistic estimation of the performance of our derived models. However, it should be highlighted that even though such validation data sets had not been used to train the models they were still derived from the same study population. Also, ideally, tuning would have been carried out on an independent, fourth portion of the data, rather than a random subsample of a single training data set (i.e. a sixth of the total data set). The limited size of our sample precluded this. Whilst cross-fold validation is also commonly used to initially tune the hyperparameters used by a machine learning approach this is also a resampling technique and would not have avoided this issue. This issue could also have contributed to some degree of overfitting, and again, adversely affected the generalisability of the model to completely independent data sets. Thus, it would be important, as part of future validation work, to assess the performance of this machine learning model in a totally independent sample, drawn from a completely separate population of patients.

Our use of a 'k-nearest' algorithm (SMOTE) to generate synthetic data to rebalance the outcome variable will have reduced some of the risk of CatBoost focussing on predicting the most common, 'negative' cases, at the expense of positive cases, where deterioration occurred. However, we only used the default value (k=5) for the 'nearest neighbours' method to generate synthetic observations for this step of our methods. Different values for k are unlikely to have substantially impacted on our findings. However, a sensitivity analysis over plausible values of k could have been performed to assess the stability of this assumption

The models derived with our machine learning approach would require the availability of reasonable amounts of computational power to be applied clinically and this represents a potential barrier for implementation into clinical practice. A simplified version of the model,

which used fewer important predictor variables, as identified via the 'importance' metric, could be used. Such a reduced model may be easier to implement in a clinical setting although may not perform as well. Thus, there would be a trade-off between model complexity and potential model performance and utility.

Ideally, all predictor variables would be completely independent of each other. However, was not the case in this data set, as with the data used in the original study. However, given the way that variables are randomly sampled in the machine learning model building process, when recombined, the less powerful predictor of a pair of variables that were relatively dependent on each other would be discarded. In the present study, this might have also been reflected in the 'importance' values for the different predictors included. Excessive dependency between predictor variables would also have caused convergence issues with our models, and such problems were not observed. Nevertheless, future studies in this area would ideally consider collecting (or combining) variables to ensure the relative independence of the predictors from each other.

This study used a representative data set drawn from a population of patients presenting to the emergency department with traumatic head injury. However, the training data were drawn only from three hospitals and therefore the generalisability cannot be assumed. Nevertheless, our use of iterative model building provided a better estimate of the uncertainty of our results, and thus the potential generalisability than would normally be reported in machine learning based predictive studies. Also, by using a recalibration model within the process we were able to change the decision threshold to increase the NPV, whilst maintaining a relatively low, but potentially acceptable PPV. Moreover, we used the latest algorithms to make the most of categorical data, as well as employing methods to adjust for relatively uncommon (unbalanced) outcomes and missing data. In common with other machine learning methods, interpreting the predictive models is much more challenging than classical approaches, although importance metrics aid somewhat in this regard.

Implications

On the basis of these findings there would not be a strong case for moving to a more complex modelling approach compared to logistic regression or rule-based algorithms at

this time. However, it may be, as more data become available, the advantages for machine learning approaches may outweigh their limitations. Also, as more data is routinely electronically captured it might be that machine learning systems are able to capitalise on a wider range of predictor variables. Certainly, to date, the situations where machine learning seems to provide an advantage over conventional statistical approaches are where there are large quantities of unstructured data to learn from. Such clinical scenarios include classification tasks related to medical imaging [18] or the natural language processing of free-text health records.[19] Such research should be reported, transparently and according to consistent reporting standards, such as those that build on the TRIPOD guidelines for prognostic studies.[20]

Our machine learning models would select patients for discharge with around a 1 in 26 chance of subsequently deteriorating. Whether this would be perceived as a clinically acceptable risk would depend on both clinicians' and patients' risk appetites and the circumstances to which a patient was being discharged to. This is likely to be seen as too high a risk if a patient is being discharged somewhere where they are not going to be monitored by family or cannot easily return to hospital if their condition changes. Moreover, current NICE guidelines advise, following head trauma, a patient should only be discharged from the ED if they can be observed at home by a responsible adult for at least 24 hours.

Future research should focus on comparing the model performance of this machine learning-based algorithm to the earlier logistic regression-based predictive model in an external validation data set. Moreover, it is important to assess the actual, real world impact of any predictive decision-making tool on actual patient care and clinical outcomes. The acceptable risk of deterioration to both patients and clinicians when discharging a patient from the ED is subjective and will vary depending on the individuals' risk appetite. Further research is needed to quantify acceptable risk of deterioration in this TBI population and how different risk prediction models could be used to support shared decision making in this context.

Conclusion

The predictive performance of our machine learning approach was similar to that of our logistic regression-based model. The risk of deterioration in a patient recommended for discharge, though relatively small, may, nevertheless, be still too high to be used clinically. Further research should be focused on developing models that provide clear advantages over existing, classical techniques for predicting outcomes in both this and external patient data sets. In addition, as in the present study, care should be taken to communicate the uncertainty over the results in order to convey a realistic appraisal of how such models are likely to perform across settings. Such rigour is essential if machine learning is to find its correct place within healthcare technology.

Author Disclosure Statement:

No competing financial interests exist.

FEL is supported by the European Union Framework 7 Collaborative European Neurotrauma Effectiveness Research in Traumatic Brain Injury ((EC grant 602150)) and NHS Trusts "Trauma Audit and Research Network - www.tarn.ac.uk" (Grant Number Not Applicable/NA).

CM is a National Institute for Health Research (NIHR) Clinical Lecturer in Emergency Medicine (Grant Number Not Applicable/NA). PAT is funded in his research by an NIHR Career Development Fellowship (CDF-2015-08-11). This publication presents independent research funded by the National Institute for Health Research, University of Sheffield, and University of York. The views expressed are those of the author(s) and not necessarily those of the University of Sheffield, University of York, the NHS, the NIHR or the Department of Health and Social Care

Authors' contributions:

The idea for the study was conceived by CM and PAT with help from FEL and LWP. The analysis was completed by PAT and LWP with clinical specialist advice regarding the interpretation of results from CM and FEL. All authors read and approved the final manuscript.

Figures:

Figure 1: Flowchart machine learning model building and validation process

Figure 2: Population Selection

References

1. NICE. CG 176 Head Injury Triage, assessment, investigation and early management of head injury in children, young people and adult. In: Department of Health, ed., 2014.
2. Miller JD. Minor, moderate and severe head injury. *Neurosurgical Review* 1986;9(1-2):135-39.
3. Haydel MJ, Preston CA, Mills TJ, et al. Indications for computed tomography in patients with minor head injury. *New England Journal of Medicine* 2000;343(2):100-05.
4. Marincowitz C, Lecky FE, Townend W, et al. The risk of deterioration in GCS13–15 patients with traumatic brain injury identified by computed tomography imaging: a systematic review and meta-analysis. *Journal of Neurotrauma* 2018;35(5):703-18.
5. Thomas BW, Mejia VA, Maxwell RA, et al. Scheduled repeat CT scanning for traumatic brain injury remains important in assessing head injury progression. *Journal of the American College of Surgeons* 2010;210(5):824-30.
6. Joseph B, Friese RS, Sadoun M, et al. The BIG (brain injury guidelines) project: defining the management of traumatic brain injury by acute care surgeons. *Journal of Trauma and Acute Care Surgery* 2014;76(4):965-69.
7. Marincowitz C, Lecky FE, Townend W, et al. A protocol for the development of a prediction model in mild traumatic brain injury with CT scan abnormality: which patients are safe for discharge? *Diagnostic and Prognostic Research* 2018;2(1):6.
8. Marincowitz C, Lecky FE, Allgar V, et al. Development of a Clinical Decision Rule for the Early Safe Discharge of Patients with Mild Traumatic Brain Injury and Findings on Computed Tomography Brain Scan: A Retrospective Cohort Study. *Journal of Neurotrauma* 2020;37(2):324-33.
9. Christodoulou E, Jie M, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* 2019
10. Collins GS, Moons KG. Reporting of artificial intelligence prediction models. *The Lancet* 2019;393(10181):1577-79.
11. Honaker J, King G, Blackwell M. Amelia II: A program for missing data. *Journal of Statistical Software* 2011;45(7):1-47.
12. CatBoost. CatBoost [14th April 2020]. Available from: https://catboost.ai/.
13. R: A Language and Environment for Statistical Computing [program]: R Foundation for Statistical Computing, 2019.
14. Kaggle Forum. Ranking of Kaggle algorithms by competitions won 2016 [Available from: https://www.kaggle.com/general/25913 accessed 20th August 2018 2018.
15. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. 2002;16:321-57.
16. Niculescu-Mizil A, Caruana R. Obtaining Calibrated Probabilities from Boosting. UAI, 2005:413-.
17. Tiffin P, Ashton H, Marsh R, et al. Pharmacokinetic and pharmacodynamic responses to caffeine in poor and normal sleepers. *Psychopharmacology* 1995;121(4):494-502.
18. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* 2018 doi: 10.1038/s41591-018-0107-6
19. Shah AD, Bailey E, Williams T, et al. Natural language processing for disease phenotyping in UK primary care records for research: a pilot study in myocardial infarction and death. *Journal of Biomedical Semantics* 2019;10(1):20. doi: 10.1186/s13326-019-0214-4
20. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. *BMC Medicine* 2018;16(1):120. doi: 10.1186/s12916-018-1099-2

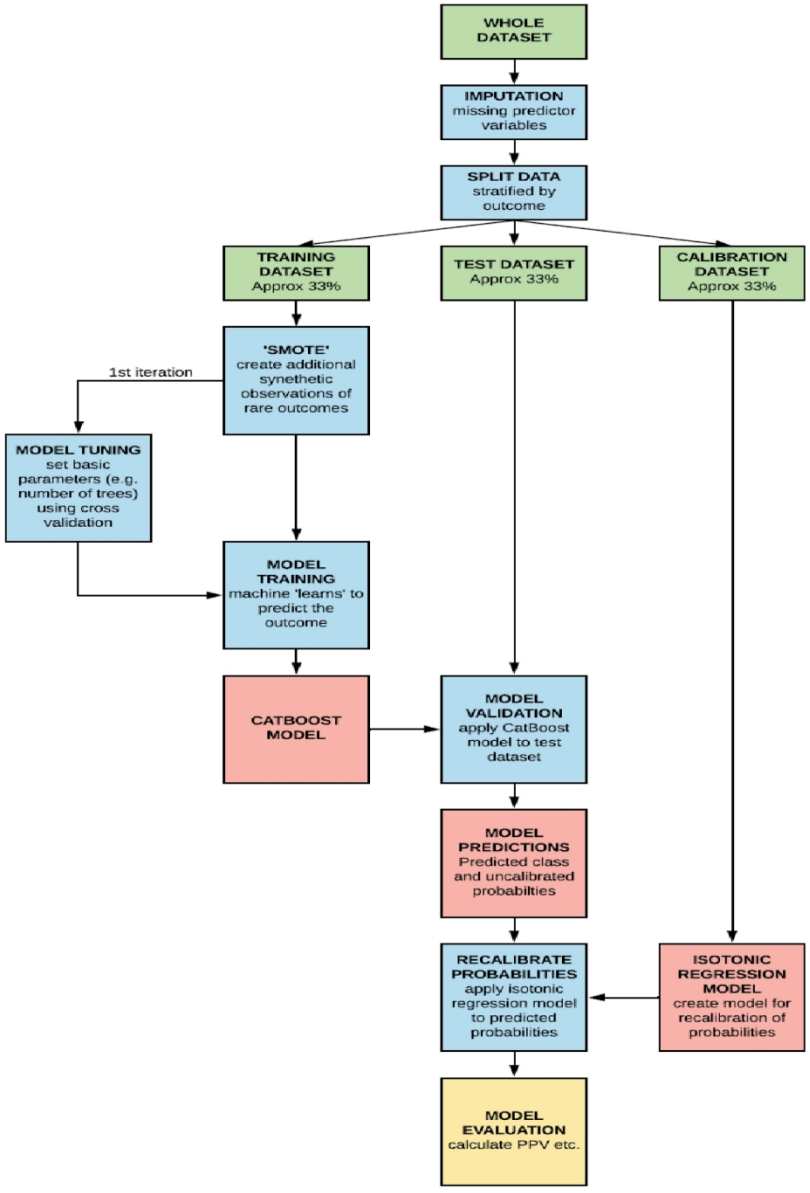

Figure 1: Flowchart machine learning model building and validation process

215x279mm (300 x 300 DPI)

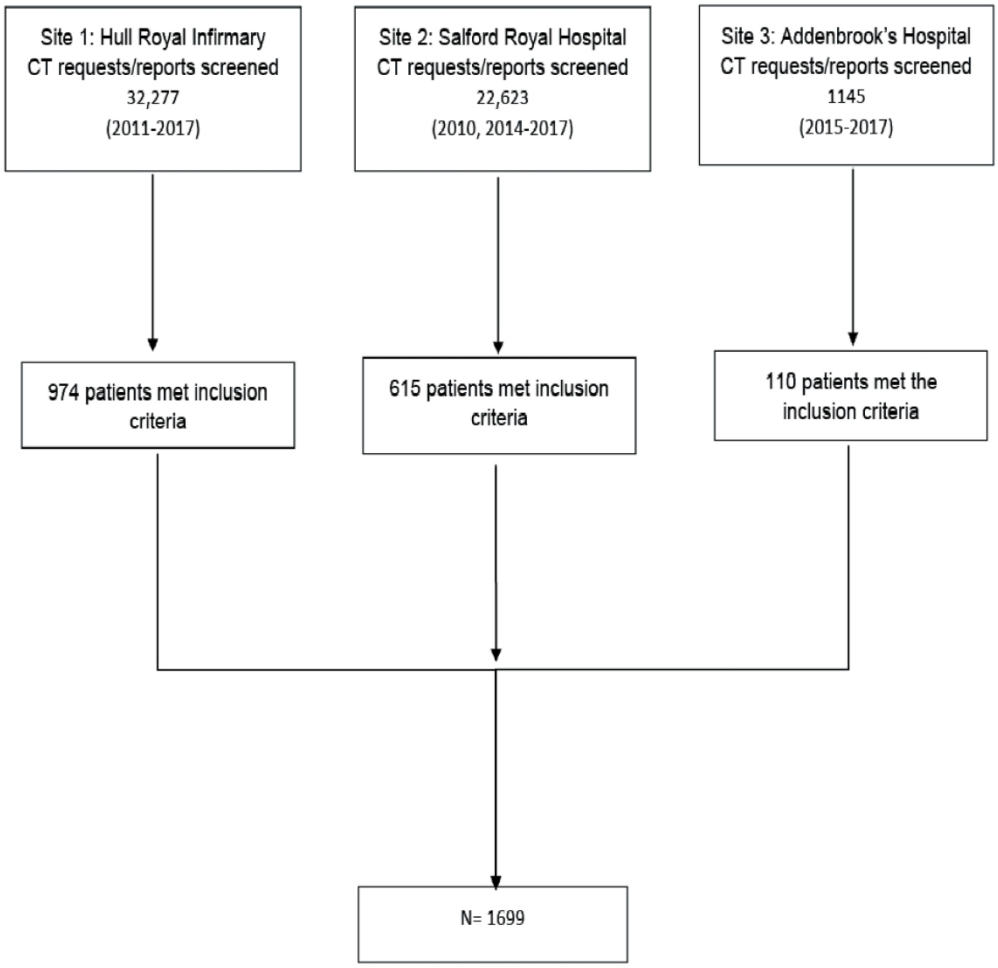

Figure 2: Population Selection

212x217mm (300 x 300 DPI)

Supplementary Material 1: Original predictive model and Clinical Risk Score

Factor	**Coefficient (optimism adjusted)**	**Risk Score Value**
Preinjury Anti-coagulation or anti-platelets	0.3	1
GCS 15 14 13	 0 (Vs) 0.4 0.7	 **GCS 15** 0 **GCS 14** 1 **GCS 13** 2
Normal first Neurological Examination	0.45	**Abnormal** 1.5
Number of Injuries on CT **1** **2** **3** **4** **5** **Diffuse**	 0 (Vs) 0.25 0.4 0.8 0.9 0.3	 **1** 0 **2** 1 **3** 1 **4** 3 **5** 3 **Diffuse** 1
Injury severity on CT* **1** simple skull fracture **2** complex Skull Fracture **3** 1-2 bleeds < 5mm **4** No or minimal mass effect **5** Significant midline shift	 0 (Vs) 0.3 0.08 0.7 1.7	 **1** 0 **2** 1 **3** 0 **4** 2 **5** 5

6 High/mixed-density lesion **7** Cerebellar/Brain stem injury	2.7 1.7	**6** 9 **7** 5
ISS (body regions excluding head)	0.2	**Up to 2 non-significant extra-cranial injuries**** 0 **Any significant extra-cranial injury or 3 or more injuries** 2
Hb	-0.01	Not included in risk score
Constant	-1.38	

*TBI severity categories are described in detail in Supplementary material 2

Supplementary Material 2: Categorisation of TBI severity

Category	Injury Description written CT report	AIS Codes	Equivalent Marshal Classification (Lesko et at[11])
1	Vault skull fractures	150000, 150400 150402	
2	Basal, depressed, open skull fractures	150200, 150204, 150205, 150206, 150404, 150406, 150408	I
3	1-2 Bleeds* /contusions total diameter <5mm	140605, 140631, 140639, 140651, 140693, 140694 (and written CT report indicated injury <5mm)	
4	Bleed/contusion No or minor mass effect	140602,140604,140606,140612,140614,140611,140620,140622, 140628,140629,140630,140632,140634,140638,140640,140642, 140644,140646,140650,140652,140654,140684,140688, 140686, 140699, 140676, 140678, 140680, 140682, 140799	II
5**	Bleed/contusion Significant midline shift or mass effect indicated in CT report	140202, 140660, 140662, 140664, 140666	III/IV
6		140608,140610,140616,140618,140624,140626,140636,140648, 140656, 140637, 140655	VI
7	Cerebellar/brainstem injury	140204,140206,140208,140210,140212,140214,140218,140299, 140402,140403,140404,140405,140406,140410,140414,140418, 140422,140426,140430,140434,140438,140442,140446,140450, 140458,140462,140466,140470,140474,140499,	VII

*Bleeds refers to subdural, extradural, intracerebral and subarachnoid haemorrhage

**Written CT reports did not allow easy differentiation in the extent of mass effect, and therefore Marshall III and IV categories were collapsed into 1 category.